IEOR 6711: Stochastic Models I

Professor Whitt, Thursday, November 1, 2012

The M/G/1 Queue

We discussed the $M/G/1$ queue; see Example 4.1 (A), p. 164, Example 4.3 (A), pp. 177-179, and Exercise 4.15 in Ross. See Examples 4.1 (B) and 4.3 (B) for a treatment of the $G/M/1$ queue. The $M/G/1$ queue has a Poisson arrival process (the M for Markov), IID service times with a general cdf G, one server and unlimited waiting space.

A Markov chain and the main results.

To get a Markov process from the queue-length process, where the queue length here is interpreted as the number of the customers in the system, including the one in service, if any, we can look at an *embedded sequence*. Specifically, for $M/G/1$, we look at the queue length just after each departure (at the departure epoch, but not counting that departing customer), while for $G/M/1$ we look at the queue length just before an arrival (at an arrival epoch, but not counting that arrival). In both cases, that leads to discrete-time Markov chains with infinitely many states. It is easy to see that the MC is irreducible and aperiodic. You apply Theorem 4.3.3 to deduce that it suffices to solve the equation $\pi = \pi P$ for a probability vector π in order to justify the existence of a limiting distribution which is also a stationary distribution. So you set about to solve $\pi = \pi P$. But that requires solving *an infinite system of linear equations*.

The first step is to construct the DTMC. Let X_n be the number of customers in the system immediately after the n^{th} departure. Let λ be the arrival rate of the Poisson arrival process. Let G be the cdf and g be the pdf of one of the i.i.d. service times, each distributed as the random variable S with mean $E[S]$. Let $c_s^2 \equiv Var(S)/(E[S])^2$ be the *squared coefficient of variation* of a service time. Let $\Rightarrow$ denote convergence in distribution.

Theorem 0.1 (*Pollaczek-Khintchine formulas for the $M/G/1$ queue*) *The stochastic process $\{X_n : n \geq 0\}$ is an irreducible aperiodic DTMC. If $\rho \equiv \lambda E[S] < 1$, then the DTMC is positive recurrent and*

$$X_n \Rightarrow X_\infty \quad as \quad n \to \infty,$$

where the distribution of X_∞ is characterized by the ***Pollaczek-Khintchine transform***

$$\hat{\pi}(z) \equiv E[z^{X_\infty}] \equiv \frac{(1-\rho)(z-1)\hat{g}(\lambda(1-z))}{z - \hat{g}(\lambda(1-z))} \ ,$$

where

$$\hat{g}(s) \equiv \int_0^\infty e^{-st}\, dG(t) \equiv \int_0^\infty e^{-st} g(t)\, dt \ .$$

is the Laplace transform of the service-time pdf g. This limiting distribution has the relatively simple ***Pollaczek-Khintchine formula*** *for its mean; i.e.,*

$$E[X_\infty] = \rho + \frac{\rho^2}{(1-\rho)} \frac{(c_s^2+1)}{2} \ .$$

Step 1 of the Proof: A Recursive Relation.

The key step is to develop a recursive relation for the random variables X_n. The key relation is the recursive expression (4.1.2) on p. 164, giving

$$X_{n+1} = X_n + Y_n - 1 \quad \text{if} \quad X_n \geq 1 \tag{1}$$

and

$$X_{n+1} = Y_n \quad \text{if} \quad X_n = 0, \tag{2}$$

where Y_n is the number of arrivals during one random service time. If $X_n \geq 1$, then a customer is in the system to enter service immediately, so we count the number of arrivals from the departure epoch until the next departure epoch, and we subtract 1 because one customer departs at the end of the interval. Equation (2) may perhaps be hard to understand. When $X_n = 0$, there is nobody in the system at the departure epoch, so we must wait an interarrival time until the first customer arrives and enters service. Hence we should add 1 for this first arrival. In this case, we get $X_{n+1} = 1 + Y_n - 1 = Y_n$, because $X_n = 0$. That explains (4.1.2) in Ross.

Step 2 of the Proof: Transition Probabilities.

Since the Markov chain is specified by the transition probabilities, we need to derive the transition probabilities. Conditioning on the length of the service time and unconditioning, we get

$$a_j \equiv P(Y_n = j) = \int_0^\infty \frac{e^{-\lambda x}(\lambda x)^j}{j!} g(x)\, dx \tag{3}$$

as in (4.1.3). We then get the transition probabilities given on p. 164.

Step 3 of the Proof: Solving $\pi = \pi P$ with generating Functions.

The next step is to solve $\pi = \pi P$. That is Example 4.3 A on pp. 177-179. For $M/G/1$, we can reduce that to a single equation by applying *probability generating functions*, as shown by Ross (and me in class). We solve for

$$\hat{\pi}(z) \equiv \sum_{n=0}^{\infty} \pi_n z^n \equiv E[z^{X_\infty}].$$

We use L'Hospital once to get π_0. In that step, we show that a proper steady-state distribution exists (there is a proper solution of the equation $\pi = \pi P$) if and only if $\rho < 1$, where $\rho = \lambda/\mu = \lambda E[S]$, with λ being the arrival rate, μ being the service rate and $\mu^{-1} \equiv E[S]$ being a mean service time, where S here denotes a generic service-time random variable. From that analysis, we see that $\pi_0 = 1 - \rho$. We get an expression for the generating function of the steady-state distribution π in terms of the arrival rate, the service rate and the Laplace transform of the service-time distribution. It is significant that $\hat{a}(z) = \hat{g}(\lambda(1 - z))$, where $\hat{g}(s)$ is the Laplace transform of a service time, and $\hat{a}(z) \equiv \sum_{j=0}^{\infty} a_j z^j$ is the generating function of the sequence $\{a_j : j \geq 0\}$, which is the distribution of the number of arrivals during a service time. (Ross uses the notation $A(s)$ instead of $\hat{a}(z)$.) In particular, a_j is the probability that the number of arrivals during a service time is j and

$$\hat{g}(s) \equiv \int_0^\infty e^{-st}\, dG(t) \equiv \int_0^\infty e^{-st} g(t)\, dt\ .$$

is the Laplace transform of the service-time pdf g. The final formula is given on page 179:

$$\hat{\pi}(z) = \frac{(1 - \rho)(z - 1)\hat{a}(z)}{z - \hat{a}(z)}\ . \tag{4}$$

The $M/M/1$ Special Case.

We can check this in simple special cases. When the service time is exponential with mean $1/\mu$,

$$\hat{g}(s) = \frac{\mu}{\mu + s} \ .$$

When this is substituted into the expression for $\hat{\pi}(z)$, we get the known geometric steady-state distribution for the $M/M/1$ queue:

$$\hat{\pi}(z) = \frac{(1-\rho)}{1-\rho z} \ .$$

Step 4 of the Proof: Deriving the Mean and Higher Moments.

For the general $M/G/1$ queue, we can differentiate the generating function $\hat{\pi}(z)$ in (4) to find the moments of the steady-state number in system. We need to differentiate and then let $z \uparrow 1$. However, when we take this limit, we get $0/0$. Hence, we need to apply L'Hospital's rule. In fact we need to use it two times to get the derivative $\hat{\pi}'(1)$, which yields the mean:

$$E[X_\infty] = \rho + \frac{\rho^2}{(1-\rho)} \frac{(c_s^2+1)}{2} \ . \tag{5}$$

These formulas are called the Pollaczek-Khintchine transform and the Pollaczek-Khintchine equation, respectively, in recognition of the fundamental work by Félix Pollaczek (around 1930). They are classic results in queueing theory.

Numerical Inversion.

The steady-state distribution is characterized by a transform, specifically a generating function, which itself depends on the transform (Laplace transform) of the service-time distribution. Numerical inversion can be used to calculate numerical values. For the $M/G/1$ queue, this application of numerical transform inversion is very straightforward; e.g., see Abate, Choudhury and W (1999). There is also a short paper on inverting generating functions, Abate and W (1992).

Pollaczek also derived (more complicated) transform solutions for the more general $GI/G/1$ queue. Our 1993 paper, Abate, Choudhury and W (1993) - available on line - shows that inversion can again be applied (specifically, for the steady-state waiting time), so that it is possible to numerically solve for the steady-state distribution (of the waiting time) of any $GI/G/1$ queue, provided that we know the transforms of the interarrival-time and service-time distributions. (The queue length can be calculated too; see Chapter X of Asmussen (2003). The general $GI/G/1$ case is much more complicated, however, because the Laplace transform of the steady-state waiting time is expressed as a contour integral. But we can perform a numerical integration to calculate the required transform values from the contour integral to use in the numerical inversion. The paper describes how to do that step.

Matrix Generalizations.

The books by Neuts (1981, 1989) discuss far-reaching matrix generalizations of the $M/G/1$ and $GI/M/1$ Markov chains. These are Markov chains that have the same kind of structure as these particular queueing Markov chains, where the elements are replaced by entire matrices. See Section XI.3 of Asmussen (2003) for a quick treatment.

Heavy-Traffic Limit and Approximation.

The Pollaczek-Khintchine transform for the steady-state number in system derived in class can be used to derive simple heavy-traffic approximations for the steady-state distribution. I post a few pages from the textbook by Gendenko and Kovalenko (1968) that do this directly. We regard the steady-state number in system at departure epochs a function of the traffic intensity ρ, and thus write $X_\infty(\rho)$. If we multiply the steady-state queue length by $(1-\rho)$, then we can show that it converges in law to an exponential distribution, with a mean that is equal to the limit of $(1-\rho)$ times the mean, which has an explicit formula:

$$(1-\rho)E[X_\infty(\rho)] \to \frac{c_s^2+1}{2} \ .$$

(Note that the limit above for the mean is elementary, given the explicit formula above.) The more general convergence in distribution is a variant of a classical theorem by Kingman (1961,1962).

Theorem 0.2 (*HT limit for the steady-state distribution in the $M/G/1$ queue*) *In the setting of Theorem* 0.1*, if $\rho \uparrow 1$, then*

$$(1-\rho)X_\infty(\rho) \Rightarrow L,$$

where

$$P(L > x) = e^{-2x/(c_s^2+1)}, \quad x \geq 0 \ .$$

The main benefit is that we see that the distribution itself is approximately exponential under heavy loading. This limit yields useful approximations for typical moderate loading. The limit is proved by taking *limits of characteristic functions*, exploiting *Taylor's series*, just as we did for the WLLN and CLT earlier. (See previous notes.)

This material from Gnedenko and Kovalenko (1968) follows the original reasoning of Kingman (1961, 1962) for establishing heavy-traffic limits. Subsequent work develops stochastic-process limits, in which the entire queue-length process is approximated by (converges to after appropriate scaling) *reflected Brownian motion*. See Whitt (2002) for an overview of this approach. Chapter 5 there provides an introduction, while Chapter 9 provides a detailed treatment of the single-server queue.

Details on the Proof of the Heavy-Traffic Limit.

Start with the generating function of the stationary distribution of the queue length after departures, as a function of the traffic intensity ρ. Let X be the random variable with this distribution. (X is the same as $X_\infty(\rho)$ above.) From (4), its generating function is

$$\hat{\pi}(z) \equiv E[z^X] \equiv \sum_{i=0}^{\infty} z^i \pi_i \equiv \sum_{i=0}^{\infty} z^i P(X=i) = \frac{(1-\rho)(z-1)\hat{g}(\lambda(1-z))}{z-\hat{g}(\lambda(1-z))}, \tag{6}$$

where $\hat{g}(s)$ is the Laplace transform of the service-time pdf, defined above. Then $\hat{g}(\lambda(1-z))$ is the generating function of the number of arrivals during a random service time, also developed above. You can multiply above and below by -1 and equivalently write

$$\hat{\pi}(z) = \frac{(1-\rho)(1-z)\hat{g}(\lambda(1-z))}{\hat{g}(\lambda(1-z))-z}. \tag{7}$$

Next convert the generating function to a characteristic function by replacing z by e^{it}. Using (7), this gives

$$\phi(t) \equiv E[e^{itX}] = \hat{\pi}(e^{it}) = \frac{(1-\rho)(1-e^{it})\hat{g}(\lambda(1-e^{it}))}{\hat{g}(\lambda(1-e^{it}))-e^{it}}. \tag{8}$$

The heavy traffic limit is for the *scaled* random variable ϵX_ϵ, where $\epsilon = 1 - \rho$, where $\rho \equiv \lambda E[S]$ is the traffic intensity, with λ being the arrival rate and $E[S]$ being the mean service time. We let $\rho \uparrow 1$, which is equivalent to $\epsilon \downarrow 0$. We do so by letting $\lambda \uparrow 1/E[S]$, so that we only change the arrival rate. The service distribution is held fixed. (It is natural to just fix $E[S] = 1$ at the outset, for simplicity. That amounts to choosing the measuring units for time.)

We are interested in the limiting behavior of the characteristic function

$$\phi_\epsilon(t) \equiv E[e^{it\epsilon X_\epsilon}] = \hat{\pi}_\epsilon(e^{i\epsilon t}) = \frac{\epsilon(1 - e^{i\epsilon t})\hat{g}(\lambda_\epsilon(1 - e^{i\epsilon t}))}{\hat{g}(\lambda_\epsilon(1 - e^{i\epsilon t})) - e^{i\epsilon t}}, \tag{9}$$

where λ is changing with ϵ, i.e., $\lambda_\epsilon \equiv (1-\epsilon)/E[S]$. That is, we have a family of systems indexed by $\epsilon \equiv 1 - \rho \equiv 1 - (\lambda E[S])$, where $\lambda \uparrow 1/E[S]$, with g and thus $E[S]$ held fixed. The family indexed by λ is equivalent to the family indexed by ϵ. The ϵ notation is used by Gendenko and Kovalenko (1968).

We now can apply the continuity theorem for characteristic functions. We will want to show that the characteristic functions (cf's) $\phi_\epsilon(t)$ converge to the cf of the exponential distribution, which is $\phi(t) = (1 - itm)^{-1}$, where the exponential random variable has mean m, as $\epsilon \downarrow 0$.

Theorem 0.3 *If $\epsilon \downarrow 0$ in the framework above, then*

$$\phi_\epsilon(t) \to \phi(t) \equiv (1 - itm)^{-1}, \tag{10}$$

where

$$m \equiv \frac{E[S^2]}{2(E[S])^2} = \frac{(c_s^2 + 1)}{2}. \tag{11}$$

It is convenient to use the parameter c_s^2 because it is a dimensionless parameter characterizing the variability (the variance except for scale). Both the traffic intensity ρ and c_s^2 are dimensionless quantities. The queue length has no time units.

Proof of the theorem. We now do the proof. To do so, we use the existence of a finite second moment to develop a Taylor series expansion for the Laplace transform $\hat{g}(s) \equiv E[e^{-sS}]$, where S is a random service time and s is a complex number; i.e.,

$$\hat{g}(s) = E[e^{-sS}] = 1 - sE[S] + s^2\frac{E[S^2]}{2} + o(|s|^2) \quad \text{as} \quad |s| \to 0. \tag{12}$$

We also use the Taylor series approximation for the exponential function

$$e^{it} = 1 + it - \frac{t^2}{2} + o(t^2) \quad \text{as} \quad t \to 0. \tag{13}$$

We thus get for the characteristic function

$$\hat{g}(\lambda_\epsilon(1 - e^{i\epsilon t})) = 1 + it\epsilon\lambda_\epsilon E[S] - t^2\epsilon^2\left(\frac{\lambda_\epsilon E[S] + (\lambda_\epsilon)^2 E[S^2]}{2}\right) + o(\epsilon^2) \quad \text{as} \quad \epsilon \to 0. \tag{14}$$

Now it only remains to insert both (13) and (14) into (9) and then let $\epsilon \downarrow 0$. We need to remember that $\rho = \lambda E[S]$ and that $\epsilon = 1 - \rho$. We first write

$$\phi_\epsilon(t) = \frac{\epsilon(1 - e^{i\epsilon t})\hat{g}(\lambda_\epsilon(1 - e^{i\epsilon t}))}{\hat{g}(\lambda_\epsilon(1 - e^{i\epsilon t})) - e^{i\epsilon t}} \tag{15}$$

We first insert the expansions to get

$$\phi_\epsilon(t) = \frac{\epsilon(-(it\epsilon) + t^2\epsilon^2 + o(\epsilon^2))(1 + it\epsilon\lambda E[S] + o(\epsilon))}{\left(1 + it\epsilon\lambda E[S] - t^2\epsilon^2\left(\frac{\lambda E[S]+\lambda^2 E[S^2]}{2}\right) + o(\epsilon^2)\right) - 1 - it\epsilon + \frac{\epsilon^2 t^2}{2} + o(\epsilon^2)} \quad (16)$$

We next add and subtract 1 in the denominator and then divide through by ϵ to get

$$\phi_\epsilon(t) = \frac{(-(it\epsilon) + t^2\epsilon^2 + o(\epsilon^2))(1 + it\epsilon\lambda E[S] + o(\epsilon))}{\left(-it(1 - \lambda E[S]) + t^2\epsilon\left(\frac{1-\lambda E[S]-\lambda^2 E[S^2]}{2}\right) + o(\epsilon)\right)} \quad (17)$$

We now simplify, recalling that $1 - \lambda E[S] = 1 - \rho = \epsilon$. Hence, equation (17) becomes

$$\begin{aligned} \phi_\epsilon(t) &= \frac{-it\epsilon + o(\epsilon)}{-it\epsilon - t^2\epsilon m + o(\epsilon)} \\ &= \frac{1 + o(1)}{1 - itm + o(1)} \to \frac{1}{1 - itm}, \end{aligned} \quad (18)$$

where

$$m \equiv \frac{E[S^2]}{2E[S]^2} \equiv \frac{c_S^2 + 1}{2}.$$

References

Abate, J., W. Whitt. 1992. Numerical inversion of probability generating functions. *Operations Research Letters*, 12 (4) 245–251.

Abate, J., G. L. Choudhury, G. L., W. Whitt. 1993. Calculation of the GI/G/1 waiting-time distribution and its cumulants from Pollaczek's formulas. *Archiv fur Elektronik und Ubertragungstechnik (AEU)*, or *International Journal of Electronics and Communication*, 47, 311-321.

Abate, J., G. L. Choudhury, G. L., W. Whitt. 1999. An introduction to numerical transform inversion and its application to probability models. in *Computational Probability*, W. Grassman (ed.), Kluwer, Boston, 257–323.

Asmussen, S. 2003. *Applied Probability and Queues*, second edition, Springer.

Gnedenko, B. V. and I. N. Kovalenko. 1968. *Introduction to Queueing Theory*, Israel Program for Scientific Translations, pages 167-169.

Kingman, J. F. C.. 1961. The single server queue in heavy traffic. *Proc. Cambridge Phil. Soc.* 57 (1961) 902-904.

Kingman, J. F. C.. 1962. On queues in heavy traffic. *J. Roy. Stat. Soc.*, Ser. B, 24 (1962) 383-392.

Neuts, M. F. 1981. *Matrix-Geometric Solutions in Stochastic Models*, The Johns Hopkins University Press.

Neuts, M. F. 1989. *Structured Stochastic Matrices of M/G/1 Type and their Applications*, Marcel Dekker.

Whitt, W. 2002. *Stochastic-Process Limits*, Springer.